Less-Known Tourist Attraction Analysis Using Clustering Geo-tagged Photographs via X-means

Jhih-Yu Lin
Graduate School of Systems Design
Tokyo Metropolitan University
Tokyo, Japan
lin-jhihyu@ed.tmu.ac.jp

Shu-Mei Wen
Department of Statistics and
Information Science in Applied
Statistics
Fu Jen Catholic University
Taipei, Taiwan
126531@mail.fju.edu.tw

Masaharu Hirota
Department of Information Science
Faculty of Informatics
Okayama University of Science
Okayama, Japan
hirota@mis.ous.ac.jp

Tetsuya Araki
Graduate School of Science and Technology
Gunma University
Gunma, Japan
tetsuya.araki@gunma-u.ac.jp

Hiroshi Ishikawa
Graduate School of Systems Design
Tokyo Metropolitan University
Tokyo, Japan
ishikawa-hiroshi@tmu.ac.jp

***Abstract*—Today, travelers can readily travel around the world using convenient transportation. Not only are opportunities to go abroad for sightseeing is increasing, but tourism industries of every country are developing indirectly. Moreover, many travelers obtain the latest tourist information from the internet for their journeys. However, most information specifically relates to popular tourist attractions, leading to crowds flocking there, which make tourists feel uncomfortable. Contrary to existing studies, which specifically emphasize analyses of popular tourist attractions, we are striving to disperse crowds from popular tourist attractions and provide more spots for travelers to choose by discovering less-known tourist attractions. This study therefore specifically examines discovery of less-known Japanese tourist attractions under the assumption that these spots exist in unfamiliar cities of tourists. According to results of analyzing geo-tagged photographs on Flickr, we use the X-means algorithm to group Japanese cities into different clusters. X-means is an extension of K-means that improved the shortcomings of K-means and which greatly reduced the probability of being trapped into a local optimum. Furthermore, these clusters were used to survey unfamiliar clusters to Japanese and Taiwanese people. Thereby, we can eliminate spots that are in familiar clusters. We propose a formula for ranking tourist attractions that lets travelers choose these spots easily. Results of verification experiments demonstrated that some less-known tourist attractions appeal to Taiwanese and Japanese. Additionally, we examined some factors that might affect respondents as they decide whether a spot is attractive to them or not.**

Keywords–Flickr; geo-tagged photograph; less-known tourist attractions；X-means

I. INTRODUCTION

In this era of the internet and smartphones, most people can readily share and record their tourist experiences on social networking services (SNSs) such as Facebook and Flickr. Numerous studies have analyzed user records of tours on SNS to elucidate user hobbies and preferences. In doing so, one can discover popular tourist attractions and recommend some tour plans for users according to their preferences [2]–[5]. Using SNSs, one can immediately obtain the newest status of friends, particularly using well-known functions related to check-in and "geo-tagged" photographs, which are useful when one wants to share a location with friends.

Aside from geolocation, diverse information is available from different SNS people users. That information includes important and useful data for research. For instance, Hausmann et al. [6] pointed out that social media contents might provide a swift and cost-efficient substitute for traditional surveys. Liu et al. [7] proposed an approach for the discovery of areas of interest (AOIs) by analyzing geo-tagged photographs and check-in information to suggest popular scenic locations and popular spots among travelers. Another study with similar aims to those of the present study used SNS users' information and geo-tagged photographs to suggest obscure sightseeing locations [8].

Most tourists receive sightseeing information through travel websites. However, almost all of these websites present well-known tourist attractions. Consequently, although the attractions are crowded and congested, visitors will be guided there. Our preliminary investigation revealed that most tourists do not like crowded spots that make them feel uncomfortable.

Many earlier studies have specifically addressed analyses of popular tourism attractions or AOIs while neglecting other unnoticed places. Our goal for this study is to improve several aspects through dispersal of crowds from more popular tourist attractions because (1) crowded popular tourist attractions make visitors feel uncomfortable, (2) foreign visitors are too numerous at popular tourist attractions, raising crime rates there, and (3) tourism to regions other than popular regions should be supported.

To accomplish our aim, we analyzed scenic geo-tagged photographs taken in Japan obtained from Flickr. After identifying some worthwhile and less-known tourist attractions, we examined them based on scenic photographs to assess their tourism value in terms of human landscapes, ecotourism, and natural landscapes. This study specifically examines natural landscapes: we used scenic photographs to appeal to travelers with natural landscapes. This study

therefore has a clearly defined research scope. Results can present more tourist attraction options for tourists and can reduce crowding at well-known tourist attractions.

Our earlier study [1] showed that over half of Taiwanese and Japanese respondents liked well-known tourist attractions and liked less-known tourist attractions. Also, questionnaire results indicated that income has little connection to travel frequency. Nevertheless, our earlier study has one point of possible improvement. Because of the influence of outliers, the grouping method used for the earlier study classified more than 70% of data into the same cluster, producing a drastically uneven data distribution. This study uses the X-means algorithm to ameliorate this shortcoming. Furthermore, we revised our earlier formula based on current questionnaire results.

The remainder of the paper is organized as follows: Section II introduces related work. Section III is an overview of the method. Section IV explains the scenic photograph evaluation method. In Section V, we present less-known tourist attraction estimation and explain our questionnaire results. Section VI presents survey questionnaire improvements and present conclusions and future work.

II. Related work

This section presents discussion of some studies related to our research, including benefits and risks of international tourism, POI and AOI, cluster analysis.

A. Benefits and Risks of International Tourism

Recently, tourism has become a development emphasis for many countries because international tourism can not only bring huge revenues; it can also have positive effects on increased long-run economic growth. Several reports have described that international tourism can bring benefits by promoting foreign exchange revenues, spurring investment in new infrastructure, stimulating other economic industries indirectly, and generating employment [9]–[14]. Moreover, Algieri et al. [15] reported that determinants of competitive advantages in tourism are important for both economically advanced and developing economies. Those determinants can help policy makers to design better strategies to strengthen activities exhibiting potential, improve performance, and enhance international competitive advantage, terms of trade, and economic growth.

Although the number of tourists continues to increase and bring huge revenues for tourism-related industries, benefits from tourism are accompanied by latent crises. Kakamu et al. [16] discovered that when the numbers of foreign visitors and the police force increase, the crime rate also increases. The rising crime rates can be expected to reduce willingness to visit and thereby tourism income [17].

B. POI and AOI

Points of interest (POIs) differ from areas of interest (AOIs). A POI is a particular spot that someone might find useful or interesting. They can be landmarks, sightseeing spots or commercial institutions of all types such as restaurants, hospitals, and supermarkets. Furthermore, POIs shown on a digital map must include some information such as name, type, longitude, and latitude. Based on data types and the discovery procedure, the approaches developed for POI are divided into two types. The first type is top-down: discovery of POI from an existing POI repository or database, such as check-in data or yellow pages that are frequently used or fit for a specific theme or target [18]–[20]. The second type is bottom-up: raw data (e.g., geotagged photos, digital footprints with implicit geographic information or metadata that involved latitude and longitude) to construct a new database or dataset that includes the POI [21]–[25].

By contrast, an AOI might include multiple geographic features or areas with no prominent landmarks, such as a café on a pedestrian street or several neighboring landmarks. Hu et al. [26] proposed that elucidating urban AOIs can provide useful information for city planners, transportation analysts, and supported location-based service providers to plan new businesses and extend existing infrastructure. After they collected Flickr photographic data of six cities in six countries, they used the DBSCAN clustering algorithm to identify urban AOI.

C. Cluster Analysis

Cluster analysis, or unsupervised classification, is one unsupervised learning technique. Cluster analysis can find objects with similar characteristics and can then group homogeneous object into clusters. Each cluster is distinct from the others. This technique is applied widely in fields such as machine learning [27]–[29], image analysis [30]–[31], information retrieval [32]–[33], bioinformatics [34]–[35], and computer graphics.

Major cluster analysis algorithms include the following.

1) Centroid-based Clustering: Centroid-based clustering is an early approach to clustering analysis in which the concept of similarity is computing the distance of a data point from the centroid of the clusters. Based on proximity, objects are assigned to clusters. Typical approaches are K-means and K-medoids. The former, K-means, is the more widely used because it has high-speed performance and easy implementation. However, K-means entails some shortcomings: K-means is sensitive to initial conditions, outliers, etc., and choosing an optimal number is difficult. According to shortcomings of K-means and our dataset, we used X-means for this study to cluster our data, as presented in Section IV.

2) Connectivity-based Clustering (Hierarchical Clustering): Clusters are constructed by calculating "distance" between objects, that can aggregate the similar object into same cluster according to the chosen similarity measure. Similarity measures include the single-linkage agglomerative algorithm, complete-linkage agglomerative algorithm, average-linkage agglomerative algorithm, centroid-linkage agglomerative algorithm, and Ward's minimum variance. In addition, hierarchical clustering is subdivided as explained below.

a) Agglomerative Approach (bottom-up): In this method, each node represents a singleton cluster from the

start. The method proceeds by agglomerating the pair of clusters of minimum dissimilarity to obtain a new cluster. Finally, nodes are merged successively based on their similarities. All nodes belong to the same cluster.

b) Divisive Approach (top-down): All nodes belong to the same cluster. The cluster is classified into sub-clusters, which are divided successively into their own sub-clusters; eventually, each node forms its own cluster. Hierarchical clustering is inappropriate processing for large amounts of data. The final result of this method is presented as a dendrogram. Clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level.

3) Density-based Clustering: Density-based clustering can identify distinctive clusters in the data by separating the contiguous region of high point density and regions of low point density. The regions of low point density are typically regarded as noise/outliers. Common examples of density models are DBSCAN [36] and OPTICS [37].

4) Grid-based Clustering: Grid-based clustering quantizes the data space into limited number of cells which form a grid structure, which can obviously reduce the computational complexity, especially for clustering very large datasets. The representative grid-based clustering algorithms are STING [38], WaveCluster [39], and CLIQUE [40].

III. OVERVIEW OF THE METHOD

Figure 1 shows that this section introduces an overview of our method. Our method comprises two components: definition of less-known tourist attractions and data construction.

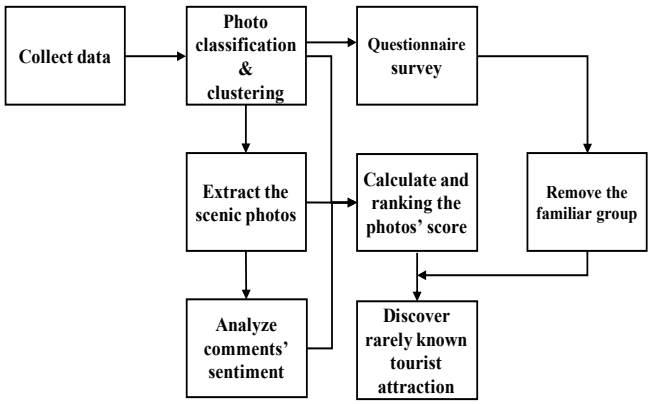


Figure 1. Overview of the method.

A. Definition of Less-Known Tourist Attractions

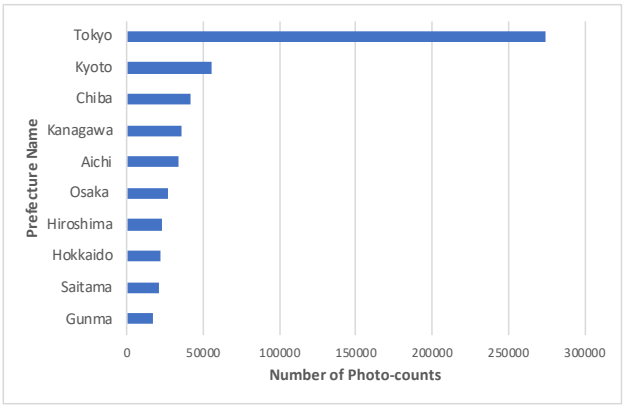
To differentiate well-known and less-known tourist attractions, we adopt two definitions of less-known tourist attractions.

Definition 1: Only some people know about this tourist attraction.

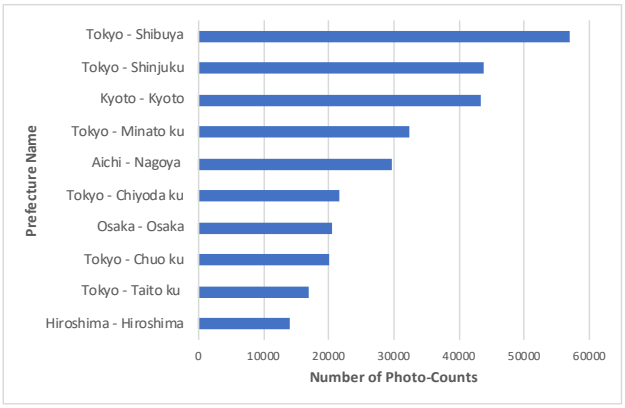
Definition 2: The tourist attraction deserves to be visited. It is attractive for tourists.

B. Data Construction

Using Flickr API, we collected 769,749 photographs taken in 2017 at geolocations throughout Japan. We extracted the photograph latitude and longitude to gather details of addresses through Google geocoding API. We found that 309 photographs were shot in the sky or on the ocean photographs had no details of addresses. We classified these photographs into different prefectures and cities according to the photograph address details. Subsequently, we calculated numbers of photographs of 47 prefectures and 1,158 cities. Figure 2 presents the Top 10 prefectures and cities in terms of the number of photographs.



(a) Top 10 Prefectures for numbers of photographs.



(b) Top 10 Cities for numbers of photographs.

Figure 2. Numbers of photographs.

Next, X-means was used to cluster prefectures and cities into different clusters to administer the questionnaire survey easily. The prefectures are divided into 4 clusters (see Table I). Prefectures are distributed into 14 clusters based on their characteristics. We also defined scores of the prefecture cluster: cluster 1 can yield 4 points, cluster 2 can yield 3 points, and so on. The city cluster score is defined according to questionnaire survey results. Furthermore, we extracted 2,671 scenic photographs with tags that mean scenic in English and Japanese (e.g., "風景", "景色", "scenery"), and collected

these photographs' comments and favorite counts. Then these photographs were ranked using formula proposed in this study. Finally, eliminating familiar city clusters according to result of questionnaire survey that is our final result.

TABLE I. CLUSTERS OF PREFECTURES

Cluster	***Prefectures***
Cluster 1	Tokyo
Cluster 2	Kyoto, Chiba, Kanagawa, Aichi
Cluster 3	Osaka, Hiroshima, Hokkaido, Saitama, Gunma, Nara, Nagano, Okinawa, Hyogo, Fukuoka
Cluster 4	Mie, Tochigi, Shizuoka, Yamanashi, Oita, Okayama, Ibaraki, Aomori, Miyagi, Gifu, Ishikawa, Wakayama, Kagawa, Niigata, Shiga, Ehime, Kumamoto, Akita, Toyama, Fukushima, Nagasaki, Yamagata, Kagoshima, Tottori, Saga, Fukui, Tokushima, Kochi, Yamaguchi, Iwate, Shimane, Miyazaki

IV. SCENIC PHOTOGRAPH EVALUATION

This section presents our approach of photograph evaluation, first illustrating how to detect the characteristic of dataset using a box plot. We can choose the most appropriate method of cluster analysis to classify our dataset. Based on the box plot result, we select X-means to cluster data in this research. We also used the elbow method to set the X-means. After analyzing the positive comments of scenic photographs by application of our formula, weight of our formula is ascertained using the entropy weight method.

A. Box Plot

Before clustering our data, we observe their characteristics. Subsequently, the suitable approach of cluster analysis can be chosen for our data. Therefore, a box plot is used to inspect the four features of Japanese cities: number of photographs of a city, the rate of the number of photographs of a city, the rate of number of photographs of a prefecture, and the average number of photographs of a prefecture.

A box plot, also called box-whisker plot, uses a statistical five number summary of dataset to visualize the data scatter. The five-number summary includes the minimum, first quartile, median, third quartile, and maximum. Moreover, this method is usually used to detect dataset outliers or to assess data symmetry.

After using feature scaling to standardize the Japanese city dataset, the box plot approach was applied to detect this dataset. Results are shown in Figure 3, which presents all cities' data scattering. A disparity between our dataset and the outliers are readily apparent in these data. These outliers cannot be eliminated because each datum represents a Japanese city. Less-known tourist attractions might exist in this city. Therefore, improper clustering methods that are sensitive to outliers should be avoided. For this study, we used X-means to cluster our data as we introduce with the next subtask.

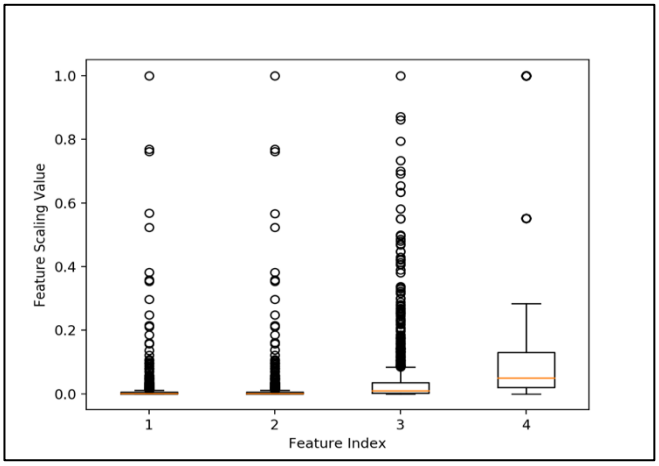


Figure 3. Box plot of dataset. Index 1 is the number of photographs of the city. Index 2 is the rate of the number of photographs of city. Index 3 is the rate of the number of photographs of the prefecture. Index 4 is the average of the number of photographs of the prefecture.

B. X-means Algorithm

The X-means algorithm is one clustering technique proposed by Pelleg and Moore [41] to improve the shortcomings of K-means. According to the BIC score, the X-means algorithm can automatically determine the optimum number of clusters that user set only minimum and maximum of clusters. Additionally, this approach greatly reduces the probability of being trapped into a local optimum and using the kd-tree to increase the computational speed.

The process steps of X-means are presented below.

1. Input dataset, setting the minimum (K_{min}) and maximum (K_{max}) parameters for the number of clusters (K).
2. Run K-means. (K= K_{min})
3. Run 2-means in each cluster according to the BIC score to decide splitting it or not.
4. If $K > K_{max}$, then stop and report the best scoring model found during the search. Otherwise, go to step 2.

Considering the outliers existing in the data, we used this method to distribute the 47 prefectures and 1,158 cities into different clusters according to their respective characteristics. Furthermore, we set the minimum cluster of X-means by referring to the result of elbow method. Finally, we obtain more scattered results than those obtained earlier by X-means. The city cluster result is presented in Table II.

TABLE II. RESULT OF CITY CLUSTER

Cluster	City counts	Percentage
Cluster 1	89	8%
Cluster 2	24	24%
Cluster 3	254	22%
Cluster 4	3	0%
Cluster 5	5	0%
Cluster 6	88	8%
Cluster 7	84	7%
Cluster 8	4	0%
Cluster 9	52	4%
Cluster 10	42	4%
Cluster 11	20	2%
Cluster 12	39	2%
Cluster 13	126	11%
Cluster 14	328	28%
Total	1,158	100 %

C. Elbow Method

The elbow method is the most popular technique used to ascertain the optimum number of clusters (*K*). The elbow method concept is calculating the total within-cluster sum of squares (wss) for each number of clusters and plotting the curve of wss. Therefore, we can ascertain the optimum number of clusters by finding the location of the warp (elbow point) in the plot of the elbow method.

Figure 4 and Figure 5 present results of the elbow method for cities and prefectures. The values of *K* are shown on the X axis. Those of wss of each *K* are shown on the Y axis. Moreover, in Figure 5, although the wss falls rapidly with *K* increasing from 1 to 4, the slope of line still has a marked change thereafter. After *K*=9, the curve goes down very slowly. Consequently, we determine optimum number of clusters as 9.

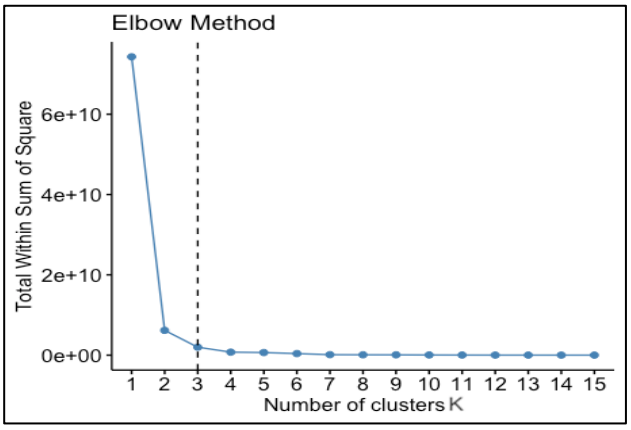


Figure 4. Elbow method of prefecture data.

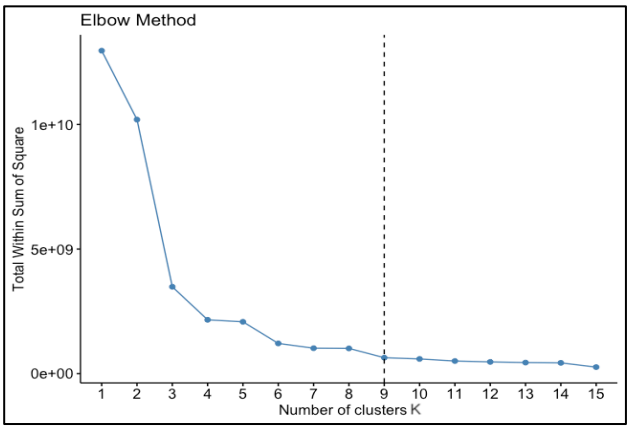


Figure 5. Elbow method of city data.

D. Comments' Sentiment

For this subtask, we assume that the positive comment is that of factors to ascertain whether this sightseeing spot is attractive for tourists to visit or not. Therefore, we collected the scenic photograph comments. Additionally, we eliminated the owner comments from the total comments because almost all of these comments are merely replies to the viewer comments. Subsequently, we analyzed these comments and extracted the positive comments, as shown in Table III.

TABLE III. NUMBER OF POSITIVE COMMENTS

	Viewer comments	Owner comments	Sum
Positive comments	1,602	248	1,850
Total comments	2,417	572	2,989

We specifically examined English and Chinese comments using TextBlob [42] and SnowNLP [43], which yielded the score of sentiment representing the probability of positive meaning. Scores of English comments' sentiments were -1 to 1. The Chinese sentiment scores were 0 to 1. To increase the accuracy of judgment, the score of English positive comments was assumed as more than 0.3; the scores of Chinese positive comments were assumed to be greater than 0.4.

E. Formula of Evaluation

Considering the definitions of less-known tourist attractions and data construction, we propose a formula to calculate the score S_i to rank the photographs.

$$S_i = \sum_{p=1}^{3} F_{pi} W_p + R_i \,, 0 < W_p < 1 \; and \; \sum_{p=1}^{3} W_p = 1 \qquad (1)$$

In equation (1), F_{1i} represents the prefecture cluster point; W_1 is F_{1i} 's weight. F_{2i} represents a city cluster point; W_2 is the weight associated with F_{2i} . F_{3i} represents the photographs' favorite counts. W_3 is F_{3i} 's weight. R_i represents the positive comment count of the photographs,

TABLE IV. PART OF JAPANESE RANKING RESULT

Address	Neighboring tourist attraction	Prefecture score	City score	Favorites	Positive comments	Score
2871 Onna, Onna-son Kunigami-gun, Okinawa, 904-0411, Japan	Resort	2	2.22	1548	56	1108.73
Yunohama hotel, 1-2-30, Yunokawacho, Hakodate-shi, Hokkaido, 042-0932, Japan	Hot spring street	2	2.02	337	13	241.83
14-16 Suehirocho, Hakodate-shi, Hokkaido, 040-0053, Japan	Kanemori Red Brick Warehouse	2	2.02	306	5	219.51
510 Tangocho Takano, Kyotango-shi, Kyoto, 627-0221, Japan	---	3	1.59	187	4	134.49
Kendou 388sen, Inuma, Kawanehon-cho Haibara-gun, Shizuoka, 428-0402, Japan	---	1	1.4	126	46	91.27
Sinkawagensi 58, Fukuoka Yatsumiya, Shiroishi-shi, Miyagi, 989-0733, Japan	---	1	1.64	123	2	88.37

TABLE V. PART OF TAIWANESE RANKING RESULT

Address	Neighboring tourist attraction	Prefecture score	City score	Favorites	Positive comments	Score
Ryuuanzi, Ryoanji Goryonoshitacho, Ukyo-ku Kyoto-shi, Kyoto, 616-8001, Japan	Temple of the Dragon at Peace	3	3.74	100	6	74.17
156 Fumoto, Fujinomiya-shi, Shizuoka, 418-0109, Japan	---	1	1.63	99	32	73.35
1070 Kodachi, Minamitsuru Gun Fujikawaguchiko Mac, Yamanashi, 401-0302, Japan	Lake Kawaguchi	1	1.89	94	19	69.5
Motosumichi, Minamigeuma-gun, Yamanashi Prefecture, Japan	---	1	1.4	92	27	68.11
Kendou60sen, Tazawako Tazawa, Semboku Shi, Akita, 014-1204, Japan	Lake Tazawa	1	1.69	69	36	51.48
86 Himata, Toyama Shi, Toyama, 930-0912, Japan	---	1	1.69	67	5	49.47

which is processed by feature scaling. In this formula, R_i is regarded as an additional score because most photographs have no associated comments. The weight of R_i is almost equal 0. The photograph favorite counts and positive comments were assumed as factors attracting someone to visit. Therefore, all scenic photographs can be ranked using this formula, as shown in Table IV and Table V.

Table IV and Table V present some Taiwanese and Japanese ranking results. The first column is the GPS address of the photograph from Google API. The second column is the neighboring popular tourist attraction. The third and fourth columns are photograph cluster scores. The fifth column shows the favorite count of photographs. The sixth column shows counts of the photograph positive comments. The last column presents the photograph score as calculated using our formula. A high score indicates that the place is attractive to travelers. In Table IV, the address of the first row is a famous resort in Okinawa. The second row presents a hotel on a famous hot spring street. The third row is a well-known tourist attraction in Hokkaido. In Table V, the first row presents a renowned and historical temple in Kyoto. The third row location is near Lake Kawaguchi: one of the Fuji Five Lakes. The place of fifth row is near Lake Tazawa, the deepest lake in Japan. Others are obscure places.

F. Entropy Weight Method (EWM)

For this study, we used EWM to set the weights used for the formula. EWM is an objective set weight method because it depends only on the discreteness of data. Actually, EWM is used widely in the fields of engineering, socioeconomic studies, etc. [44]–[46].

In information theory, entropy is a kind of uncertainty measure. When information is greater, uncertainty and entropy are smaller. Based on the entropy information properties, one can estimate the randomness of an event and the degree of randomness through calculation of the entropy value. Furthermore, entropy values are used to gauge a sort of discreteness degree of index. When the degree of discreteness is larger, the index affecting the integrated assessment is expected to be greater.

To complete the setting of the formula weights, we require the steps, as described below.

1) Calculate the ratio (P_{ij}) of the i-th index under the j-th index. Therein, x_{ij} denotes the j-th index of the i-th sample.

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, (i = 1, \dots, n;\ j = 1, \dots, m) \quad (2)$$

2) *Calculate the entropy value (e_j) of the j-th index.*

$$e_j = -k \sum_{i=1}^{n} P_{ij} \ln\left(P_{ij}\right), (j = 1, \dots, m;\ k = \frac{1}{\ln(n)} > 0) \quad (3)$$

3) *Calculate the discrepancy of information entropy (d_j).*

$$d_j = 1 - e_j, (j = 1, \dots, m) \quad (4)$$

4) *Calculate the weight (w_i) of each index.*

$$w_i = \frac{d_j}{\sum_{j=1}^{m} d_j}, (j = 1, \dots, m) \quad (5)$$

We analyzed the prefecture cluster score (F_{1i}), the city cluster score (F_{2i}), and the favorite counts (F_{3i}) of 2,671 scenic photographs. Results show the weight of the formula in this research by EWM. In the equations (1), Taiwanese and Japanese weights differ because their city clusters are assigned distinct scores based on the results of questionnaire surveys. The weight results are shown in Table VI: Taiwanese W_1 is equal to 0.1338; W_2 is equal to 0.1346 and W_3 is equal to 0.7316. Japanese W_1 is equal to 0.1619; W_2 is equal to 0.1228 and W_3 is equal to 0.7152.

TABLE VI. TAIWANESE AND JAPANESE WEIGHTS

	W_1	W_2	W_3
Taiwanese weight	0.1338	0.1346	0.7316
Japanese weight	0.1619	0.1228	0.7152

V. LESS-KNOWN TOURIST ATTRACTION ESTIMATION

A. Familiarity Level of Japanese City

For this study, we assume the less-known tourist attractions might be included in unfamiliar city clusters. Accordingly, a questionnaire was designed and administered to 115 Taiwanese and 123 Japanese people to ascertain their level of familiarity with Japanese cities. Nevertheless, surveying levels of familiarity of each city (1,158 cities) from respondents was difficult. For that reason, we clustered the Japanese city data. Thereby, we were able to select a city's name randomly from each cluster to decrease the number of questionnaire questions. It was easier to find which cities were unfamiliar to respondents.

According to the scale of each cluster, 30 city names were selected randomly for this questionnaire. Participants were provided with five choices to answer the city questions: (1) I have absolutely no idea. (2) I have heard of this city, but I do not know its tourist attractions. (3) I have heard of this city and know its tourist attractions. (4) I have been to this city, but I do not know its tourist attractions. (5) I have been to this city and know its tourist attractions. A respondent choosing option (1) is assigned 1 point for this question; option (2) yields 2 points, and so on, with higher scores representing greater familiarity with this city.

Considering that we used the survey sampling approach to conduct the questionnaire survey, it might include sampling error. To decrease inaccuracy from the sampling error, we categorized the cluster as a less-known one using *t*-tests and *p*-value.

$$\mathrm{t} = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}} \quad (6)$$

Student's t-test can determine whether a statistically significant difference exists between the means of two unrelated groups. This approach has three types: one-sample t-test, independent-sample t-test, and paired sample t-test. For this study, we used one-sample t-test to analyze our result of questionnaire survey, which can compare population means with a sample mean, and found their relation. In the following equation (6), $\overline{x}$ represents the sample means, μ denotes the population mean, s stands for the sample standard deviation, and n is the sample size. After calculating the t-test values, we used *p*-value to determine whether the sample mean was greater than the population mean or not. If the *p*-value of cluster was less than 0.05, then we judged this cluster as an unfamiliar cluster. Conversely, the cluster will be categorized into familiar clusters when the *p*-value is greater than 0.05.

TABLE VII. TAIWANESE UNFAMILIAR CITY CLUSTERS

Cluster	Sample mean	*t*-test value	*p*-value	Unfamiliar
Cluster 1	1.63	-1.40	0.08	
Cluster 2	1.89	2.05	0.98	
Cluster 3	1.40	-7.63	0.00	✓
Cluster 4	3.74	18.07	1.00	
Cluster 5	2.90	8.55	1.00	
Cluster 6	1.70	0.12	0.55	
Cluster 7	1.56	-2.50	0.01	✓
Cluster 8	2.55	6.49	1.00	
Cluster 9	1.74	0.62	0.73	
Cluster 10	1.69	-0.06	0.47	
Cluster 11	1.59	-1.11	0.13	
Cluster 12	1.52	-2.82	0.00	✓
Cluster 13	1.37	-7.22	0.00	✓
Cluster 14	1.31	-11.33	0.00	✓
Population mean	1.69	---	---	---

TABLE VIII. JAPANESE UNFAMILIAR CITY CLUSTERS

Cluster	Sample mean	*t*-test value	*p*-value	Unfamiliar
Cluster 1	2.02	0.22	0.59	
Cluster 2	2.96	7.14	1.00	
Cluster 3	1.61	-7.24	0.00	✓
Cluster 4	4.47	29.79	1.00	
Cluster 5	3.51	10.23	1.00	
Cluster 6	1.80	-3.53	0.00	✓
Cluster 7	1.47	-11.01	0.00	✓
Cluster 8	3.24	9.55	1.00	
Cluster 9	2.32	3.46	1.00	
Cluster 10	2.22	2.28	0.99	
Cluster 11	1.59	-4.42	0.00	✓
Cluster 12	2.09	0.88	0.81	
Cluster 13	1.64	-6.50	0.00	✓
Cluster 14	1.40	-15.37	0.00	✓
Population mean	2.01	---	---	---

Table VII and Table VIII show that we calculated the average scores of respective clusters. Table VII and Table VIII present results of unfamiliar clusters. The first column shows the number of clusters. The second column is each cluster average score from the questionnaire survey. The third column shows the t-test statistic value. The fourth column presents p-values. The last column presents which cluster is unfamiliar. In Table VII, one can understand that cluster 3, cluster 7, cluster 12, cluster 13, and cluster 14 are unfamiliar to Taiwanese. Moreover, Table VIII shows that Japanese people are unfamiliar with cluster 3, cluster 6, cluster 7, cluster 11, cluster 13, and cluster 14. Finally, we can remove these familiar clusters from the ranking results of Section IV as our aim.

B. *Verification Experiment*

Based on the discussion presented above, we can ascertain which group is unfamiliar to the Taiwanese respondents (clusters 3, 7, 12–14) and to the Japanese respondents (clusters 3, 6, 7, 11, 13, 14). In this section, we also use the questionnaire to verify these less-known tourist attractions, which are obscure but attractive to respondents.

For the verification experiment, we extracted the top 10 less-known tourist attractions from nine cities of the Taiwanese and Japanese unfamiliar clusters to investigate 10 Taiwanese people (who have touristic experience in Japan) and 10 Japanese people, whose questionnaires responses were dissimilar. Two questions were asked for each attraction: "Do you know this city?" If respondents probably know this city, then the answer was "Yes." The second question was "According to this photograph, do you want to visit this place of city?" For the second question, respondents assigned a score of 1–5 for the attraction, with a higher score indicating greater attraction.

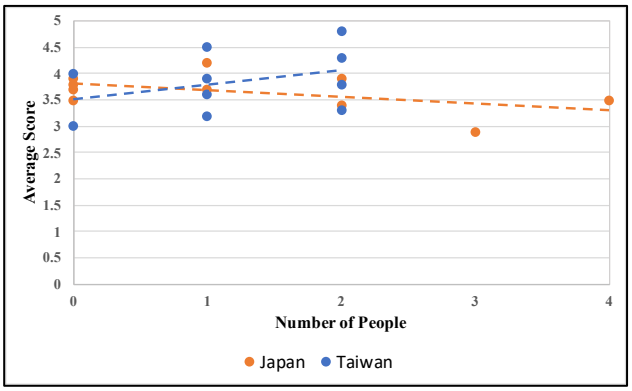


Figure 6. Result of verification experiment.

This result of the verification experiment showed that these places are known by extremely few people, which is better than the previous result. In Figure 6, each point represents a place in the questionnaire, the horizontal axis presents how many people know the less-known tourist attraction, the vertical axis shows the attractive level of each less-known tourist attraction. The Taiwanese result presents the average score of four places are over than 4 points, which means these places are attractive for Taiwanese respondents. Especially, one place score approaches the full mark. Some Taiwanese respondents reported that a few scenic photographs are similar to scenery in their own country, which might influence their decision. Moreover, for the place with the lowest score, the photograph quality was not very high. As a result, the respondents assigned few points for this place. The required cost of Taiwanese includes a monetary cost and time cost, which are higher than those of Japanese people. Consequently, Taiwanese prefer to choose tourism attractions that include local characteristics or exceptional landscapes.

For the Japanese result, although only one spot yielded over 4 points, two other spots yielded over 3.5 points, indicating that Japanese respondents are not excluded from visiting these spots. Furthermore, we investigated the answers of each Japanese respondent in depth and detected that the disparity between their decisions decreased the average. This situation expresses that respondents chose the answer according to their preference of scenic spots. For instance, someone who likes the ocean, but does not like mountains will assign a higher score for seascape photographs. Therefore, the average of each spot is less than 4 points. Figure 6 also shows an interesting situation. The Taiwanese trendline shows that the number of people and average score are in direct proportion, but the Japanese trendline is inverse to that of the Taiwanese curve: Japanese people prefer to visit less-known tourist attractions.

VI. DISCUSSION AND CONCLUSION

We proposed a novel method to identify less-known tourist attractions for people of different nationalities. By collecting and analyzing Flickr photograph information, we classified them into prefectures and cities. Subsequently, we classified these prefectures and cities into different groups.

Additionally, we used a questionnaire to survey Taiwanese respondents and Japanese respondents. We obtained unfamiliar city clusters of Taiwanese and Japanese respondents. Scenic photographs were ranked using the formula for this research. Familiar city clusters were removed from respondent ranking results. A second questionnaire survey verified our results. Through this research, we found less-known tourist attractions for travelers.

The first questionnaire survey gave the surprising result that Taiwanese respondents are more familiar with Japanese cities than Japanese respondents are. The reason might be that Taiwan and Japan are neighboring countries. In addition, air travel from Taiwan to Japan is cheaper, which might engender a higher frequency of Taiwanese taking trips to Japan. Results show that most Taiwanese respondents prefer individual travel in Japan to travelling with groups.

The verification experiment revealed an interesting thing: we provide two seascape photographs from distinct spots for Taiwanese respondents. One photograph shows the "torii", which is the traditional gate of Japanese shrines. The other only has a clear ocean and beach. Two Taiwanese respondents said that the seascape is common in Taiwan, but they are very interested in the first seascape because of this spot, which includes "torii." Scenic photographs including some special landmarks are expected to increase the attractiveness of these spots.

Interviews of some Taiwanese respondents to ascertain what factors lead them to prefer to travel in Japan indicated four main reasons it is attractive to Taiwanese. The first reason is that air tickets are cheaper and the flight time is short. The second reason is that the Japanese environment is neat and tidy. Furthermore, public security is high. The third reason is that Japanese foods are delicious and exquisite. The fourth reason is that Japanese character and culture are similar to those of Taiwan, which can help Taiwanese people travel in Japan easily.

As future work, after collecting and analyze more photographs taken in distinct years, we expect to sort the photographs with lowest quality from our data and remove them. Providing higher-quality photographs for travelers might induce them to visit. Considering more factors of discovering less-known tourist attractions, we expect to improve the formula used for this research. Less-known tourist attractions will be classified into different types (e.g., ocean, mountain, sky), seasons, weather, days, and nights according to the times and contents of photographs. We also want to provide a personal recommendation service based on collaborative filtering.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers 19K20418.

REFERENCES

[1] J. Lin, S. Wen, M. Hirota, T. Araki, and H. Ishikawa, "Analysis of Rarely Known Tourist Attractions by Geo-tagged Photographs," MMEDIA, Mar. 2019.

[2] K. H. Lim, J. Chan, C. Leckie, and S. Karunasekera, "Personalized Tour Recommendation Based on User Interests and Points of Interest Visit Durations," IJCAI, pp. 1778–1784, Jul. 2015.

[3] I. Memon, L. Chen, A. Majid, M. Lv, I. Hussain, and G. Chen, "Travel Recommendation Using Geo-tagged Photos in Social Media for Tourist," International Journal of Wireless Personal Communications, vol. 80, pp. 1347–1362, Feb. 2015.

[4] S. Jiang, Z. Qian, T. Mei, and Y. Fu, "Personalized Travel Sequence Recommendation on Multi-Source Big Social Media," IEEE Trans. Big Data, vol. 2, no. 1, pp. 43–56, Mar. 2016.

[5] X. Peng and Z. Huang, "A Novel Popular Tourist Attraction Discovering Approach Based on Geo-Tagged Social Media Big Data," ISPRS International Journal of Geo-Information, vol. 6, no. 7, pp. 216, Jul. 2017.

[6] A. Hausmann et al. "Social Media Data Can be Used to understand Tourists' Preferences for Nature-Based Experiences in Protected Areas," Conservation Letters, vol. 11, no. 1, Jan. 2017.

[7] J. Liu, Z. Huang, L. Chen, H. T. Shen, and Z. Yan, "Discovering areas of interest with geo-tagged images and check-ins," ACM Multimedia, pp. 589–598, Nov. 2012, ISBN: 978-1-4503-1089-5.

[8] C. Zhuang, Q. Ma, X. Liang, and M. Yoshikawa, "Discovering Obscure Sightseeing Spots by Analysis of Geo-tagged Social Images," ASONAM, pp. 590–595, Aug. 2015, ISBN: 978-1-4503-3854-7.

[9] S. F. Schubert, J. G. Brida, and W. A. Risso, "The impacts of international tourism demand on economic growth of small economies dependent on tourism," Tourism Management, vol. 32, pp. 377–385, Apr. 2011.

[10] A. Konstantinos, "Scale of hospitality firms and local economic development-dividence from Crete," Tourism Management, vol. 23, pp. 333–341, Aug. 2002, ISSN 0261-5177.

[11] R. R. Croes, "A paradigm shift to a new strategy for small island economies: embracing demand side economics for value enhancement and long term economic stability," Tourism Management, vol. 27, pp. 453–465, Jun. 2006.

[12] F. Michael, "Tourism as a feasible option for sustainable development in small island developing states (SIDS): Nauru as a case study," Pacific Tourism Review, vol. 3, no. 2, pp. 133–142(10) ,1999.

[13] B. Lin and H. Liu, "A study of economies of scale and economies of scope in Taiwan international tourist hotels," Asia Pacific Journal of Tourism Research, vol. 5, pp. 21–28, Apr. 2007.

[14] G.I. Crouch and J.R.B. Ritchie,"Tourism, Competitiveness, and Societal Prosperity," vol. 44, pp. 137–152, Mar. 1999.

[15] B. Algieri, A. Aquino, and M. Succurro, "International competitive advantages in tourism: An eclectic view," Tourism Management Perspectives, vol. 25, pp. 41–52, Jan. 2018.

[16] K. Kakamu, W. Polasek, and H. Wago, "Spatial interaction of crime incidents in Japan," Mathematics and Computers in Simulation, vol. 78, no. 2, pp. 276–282, Jul. 2008.

[17] D. Altindag, "Crime and International Tourism," Journal of Labor Research, vol. 35, no. 1, pp. 1–14, Mar. 2014, doi: 10.1007/s12122-014-9174-8.

[18] H. Chuang, C. Chang, T. Kao, C. Cheng, Y. Huang, and K. Cheong, "Enabling maps/location searches on mobile devices: Constructing a POI database via focused crawling and information extraction," Int. J. Geogr. Inf. Sci., vol. 30, pp. 1405–1425, Jan. 2016.

[19] D. Jonietz and A. Zipf, "A. Defining fitness-for-use for crowdsourced points of interest (POI)," ISPRS International Journal of Geo-Information, vol. 5, Aug. 2016.

[20] A. Rousell, S. Hahmann, M. Bakillah, and A. Mobasheri. "Extraction of landmarks from OpenStreetMap for use in

navigational instructions," In Proceedings of the 18th AGILE International Conference on Geographic Information Science, Jun. 2015.

[21] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring Millions of Footprints in Location Sharing Services," Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 81–88, Jun. 2011.

[22] E. Spyrou, M. Korakakis, V. Charalampidis, A. Psallas, and P Mylonas. "A Geo-Clustering Approach for the Detection of Areas-of-Interest and Their Underlying Semantics," Algorithms, Mar. 2017, DOI:10.3390/a10010035.

[23] A. Skovsgaard, D. Idlauskas, and C. S. Jensen. "A clustering approach to the discovery of points of interest from geo-tagged microblog posts," 2014 IEEE 15th International Conference on Mobile Data Management (MDM), pp. 178–188, Jul. 2014, DOI: 10.1109/MDM.2014.28.

[24] D. D Vu, H. To, W. Shin, and C. Shahabi. "GeoSocialBound: An Efficient Framework for Estimating Social POI Boundaries Using Spatio–Textual Information," Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data, Jun. 2016, DOI:10.1145/2948649.2948652.

[25] C. Kuo, T. Chan, I. Fan, and A. Zipf, "Efficient Method for POI/ROI Discovery Using Flickr Geotagged Photos," ISPRS International Journal of Geo-Information, Mar. 2018, DOI:10.3390/ijgi7030121.

[26] Y. Hu, S. Gao, K. Janowicz, B. Yu, W. Li, and S. Prasadd, "Extracting and understanding urban areas of interest using geotagged photos," Computers, Environment and Urban Systems, vol. 54, pp. 240–254, Nov. 2015.

[27] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," International Conference on Machine Learning (ICML), pp. 478–487, 2016.

[28] L. Wang, "Discovering phase transitions with unsupervised learning," Phys. Rev. B 94, Nov. 2016, DOI: 10.1103/PhysRevB.94.195105.

[29] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for Internet of Things data analysis: A survey," Digital Communications and Networks, vol. 4, pp. 161–175, Aug. 2018.

[30] N. Dhanachandra, Y. J. Chanu, and K. M. Singh, "Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm," Procedia Computer Science, vol. 5, pp. 764–771, 2015.

[31] N. B. Bahadure, A. K. Ray, and H. P. Thethi, "Performance analysis of image segmentation using watershed algorithm, fuzzy C-means of clustering algorithm and Simulink design," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), Mar. 2016.

[32] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," arXiv preprint arXiv:1707.02919, 2017.

[33] W. B. A. Karaa, A. S. Ashour, D. B. Sassi, P. Roy, N. Kausar, and N. Dey, "MEDLINE Text Mining: An Enhancement Genetic Algorithm Based Approach for Document Clustering," Applications of Intelligent Optimization in Biology and Medicine, pp. 267–287, Mar. 2015.

[34] D. Granato, J. S. Santos, G. B. Escher, B. L. Ferreira, and R. M. Maggio, "Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective," Trends in Food Science & Technology, vol. 72, pp. 83–90, Feb. 2018.

[35] M. D. M. Fernández-Arjona, J. M. Grondona, P. Granados-Durán, P. Fernández-Llebrez, and M. D. López-Ávalos "Microglia Morphological Categorization in a Rat Model of Neuroinflammation by Hierarchical Cluster and Principal Components Analysis," Front Cell Neurosci., vol. 11, Aug. 2017, DOI: 10.3389/fncel.2017.00235.

[36] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231, 1996.

[37] M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, vol. 28, pp. 49–60, 1999.

[38] W. Wang, J. Yang, and R. R. Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining," In Proceedings of the 23rd International Conference on Very Large Data Bases, pp. 186–195, Aug. 1997, ISBN:1-55860-470-7.

[39] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases," In Proceedings of the 24rd International Conference on Very Large Data Bases, pp. 428–439, Aug. 1998, ISBN:1-55860-566-5.

[40] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, pp. 94–105, Jun. 1998, ISBN:0-89791-995-5.

[41] D. Pelleg and A. W. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," In Proceedings of the 17th International Conf. on Machine Learning, pp.727–734, Jul. 2000.

[42] TextBlob.[Online]. Available from: https://pypi.org/project/textblob/ 2019.08.15

[43] SnowNLP.[Online].Available from: https://pypi.org/project/snownlp/ 2019.08.15

[44] Y. He, H. Guo, M. Jin, and P. Ren, "A linguistic entropy weight method and its application in linguistic multi-attribute group decision making," Nonlinear Dynamics, vol. 84, no. 1, pp. 399–404, Jan. 2016.

[45] Y. Ji, G. H. Huang, and W. Sun, "Risk assessment of hydropower stations through an integrated fuzzy entropy-weight multiple criteria decision making method: A case study of the Xiangxi River," Expert Systems with Applications, vol. 42, no. 12, pp. 5380–5389, Jul. 2015.

[46] A. Delgado and I. Romero, "Environmental conflict analysis using an integrated grey clustering and entropy-weight method: A case study of a mining project in Peru," Environmental Modelling & Software, vol. 77, pp. 108–121, Mar. 2016.